

Designing and Developing Person-based and Topic-Based Data Collection Instruments

Emily A. Johnson, Thomas C. Melaney, US Census Bureau

1. Summary

The American Community Survey (ACS) periodically conducts content testing in order to determine how proposed changes will affect the data collection operations. One of the recommended changes to the ACS was to modify the collection of demographic data from a person-based to a topic-based format. The ACS instrument used for Computer Assisted Telephone Interviewing (CATI) and Computer Assisted Personal Interviewing (CAPI) operations is programmed using Blaise 4.6.965. This paper examines the issues that were involved in developing the alternate ACS instruments that could be used for the collection of demographic data in person-based mode and in topic-based mode. The development issues included assuring that both of the data collection instruments conformed to the questionnaire design requirements and that both of the instruments conformed to the data processing requirements.

2. Background

The American Community Survey is a continuing survey that collects data from a sample of households throughout the United States. The data collection operation occurs each month with a different set of households. The U.S. Census Bureau collects and tabulates this data to produce current and reliable reports of the characteristics of households and household members. The ACS will replace the “long form” in future censuses and is a critical element in the Census Bureau’s reengineered 2010 Census plan. After extensive testing, the ACS reached full implementation in 2005. Currently, each month about 250,000 households receive a paper questionnaire by mail. After a month, about 100,000 households from which a paper questionnaire has not been returned are contacted by telephone to complete an interview. After another month, about 45,000 households for which the interview could not be completed by mail or telephone are visited for a personal interview.

Because of the large size of the data collection operation, and the high profile and the extensive use of the tables from the ACS, there is also an associated American Community Survey Content Test that is used to verify that proposed changes to the ACS will perform as expected without unintended consequences. In 2007, a modification was proposed which would be operationally tested in the ACS Content Test to change the data collection structure of the demographic section of the instrument. Until this time, the sections of the ACS program were in a person-based format; that is, all demographic information was collected about the first identified individual in the household, and then the next individual and so on. The proposed change was to collect the data for the demographic section in a topic-based format; that is, asking the first question in the section for each member of the household, then the next question for each member, and so on.

The question of whether the person-based or topic-based data collection method is more appropriate for certain survey topics have been tested at the U.S. Census Bureau. In 1997, a two-month test was conducted using alternate design structures of the ACS CATI instrument: person-based and topic-based. The results of this study were included in the paper “ACS/CATI Person-Based/Topic-Based Field Experiment” by Jeff Moore and Laurie Moyer of the Center for Survey Methods Research in the Statistical Research Division of the U.S. Census Bureau.

The test results examined length of interviews, response rates, refusal rates, and interviewer and respondent perceptions about the data collection process. The results generally indicated advantages for the topic-based approach. Some items were noted as needing additional research.

Following this test, the ACS retained a person-based approach pending additional review because the results for the questions collecting “income” data suggested that the person-based approach was more appropriate. In particular, the item non-response rates for the “wage/salary income” and for “total income” questions were somewhat greater using the topic-based approach.

For the 2010 Decennial Census operation, as well as the 2008 “Dress Rehearsal”, the U.S. Census Bureau decided to use a topic-based approach for the data collection operation. The questions for the Decennial Census ask about the demographic characteristics of persons within each household. When it is possible and appropriate, the ACS data collection question design should be consistent with that used for the decennial operation. Therefore, the ACS needed to be modified to use a topic-based approach for the data collection of the demographic questions for relationship, gender, date of birth and age, Hispanic origin and race.

The ACS 2007 Content Test needed to verify that moving to a topic-based design would not cause operational problems with the ACS 2008 data collection operation or with subsequent data processing operations. It was decided to use two instruments to allow a clear comparison between person-based and topic-based demographic data collection.

In the discussion that follows, please note that our example instrument is somewhat different from the actual Content Test instrument. These modifications were added to demonstrate appropriate enhancements that could have been incorporated had time permitted.

3. Approach to Design and Development

3.1 Client Request

The client was interested in conducting a test that would allow the comparison of the two data collection methods side by side. Two separate samples of respondents would be generated, with one group being asked questions in the original person-based method, and the other in the topic-based method. This required the development of two separate instruments, one using the person-based approach currently used in production for the American Community Survey, the other using the new topic-based approach. Other than differing demographic sections, the two instruments had to be kept identical. Output processing staff required that the same data be output from both instruments.

Despite the two demographic sections operating differently, the sponsor's requirement was that the two instruments have the same look and feel, with an identical table layout, and all of the person data still being displayed across a row. With the new topic-based instrument closely resembling the instrument being fielded in production, a minimal amount of training would be required. It would also lower the impact on the interviewer of changing from one content test instrument to the other.

A free-form navigation approach to collecting the data in the table would not fulfill the client's requirements, as it was not sufficiently structured. The goal was to continue to guide the interviewer from question to question; however, the client preferred to move through the table column by column, instead of row by row.

3.2 Specification Development

The two programs had a similar structure, with many modules outside of the demographic section being identical, and the naming conventions between the instruments remaining nearly the same.

The specification needed to indicate the new question flow through the topic-based demographic section. Because the order in which the questions were to be asked was not changing, the spec-writer was able to start with the current person-based specification and modify it. In a "Special Instructions" section of the field level specification, the spec-writer documented that a question was to be asked of each person in the household before moving on to the next question (see Figure 3.2.1).

Figure 3.2.1 – Specification example describing flow through grid

Special Instr.

2. Interviewers are to navigate vertically in the grid, that is, each question set must be answered for each person listed before moving to the next question set for each person in the grid. The relationship items for each respondent (RELP) will be answered for each person before moving to SEX.

Certain questions were to be answered as groups, and the specification needed to account for these exceptions to the straight topic-based approach. For example, the Hispanic Origin questions (HISA, HISB, and HISW) had to be answered as a question set, asking the leading questions (HISA), then the probe questions (HISB and HISW) if appropriate, for a single person before moving on to the next household member. Again making use of the "Special Instructions" for a specific field, the sponsor was able instruct the programmer as to whether or not the field was to be part of one of these special topic groups (see Figure 3.2.2).

Figure 3.2.2 – Specification example describing flow through grid with question set

Special Instr.

2. Interviewers are to navigate vertically in the grid, that is, each question set must be answered for each person listed before moving to the next question set for each person in the grid. The Hispanic origin items for each respondent (HISA and HISB and HISW, if applicable) will be answered for each person before moving to the race questions.

A significant reason for asking the questions in a topic-based approach was to make use of different wording that would ease respondent burden and improve interview flow. By

asking the same question of each person, the full question text would not need to be repeated, and an abbreviated form of the question could be asked for additional members in the household. Some questions could have three or four different variations, depending on how many times the question has already been asked and what question preceded it if it is part of a topic group with probe questions. Again, the specification needed to account for these significant wording changes. The specification writer created fills in the question text and based the fill text on whether the question had already been asked of a preceding household member.

In the question wording specification example below (see Figure 3.2.3), one of the more complex scenarios in the demographic section, there is question text including an introduction and example list to be read the very first time the RACE question is asked of someone in the household, along with slightly shorter text for the second time the question is asked. There follow two wording possibilities for the third or more times the question is asked – one if the previous person was asked any follow-up race questions and one if only the initial question was asked.

Figure 3.2.3 – Specification Example with Variable Wording

Question Text	
	If first time this question is asked:
	I'm going to read a list of race categories. You may choose one or more races. {Fill 1: For this survey, Hispanic origins are not races.}
	{Fill 2: Is <Name>/ Are you} White; Black, African American, or Negro; American Indian or Alaska Native; Asian; Native Hawaiian or Other Pacific Islander; or Some other race?
	<11> White <12> Black, African American, or Negro <13> American Indian or Alaska Native <14> Asian <15> Native Hawaiian or Other Pacific Islander <16> Some other race
	If second time question is asked:
	{Fill 2: Is <Name>/ Are you} White; Black, African American, or Negro; American Indian or Alaska Native; Asian; Native Hawaiian or Other Pacific Islander; or Some other race?
	<11> White <12> Black, African American, or Negro <13> American Indian or Alaska Native <14> Asian <15> Native Hawaiian or Other Pacific Islander <16> Some other race
	If not first or second time question is asked and previous person RACT=13, 14, 15 or 16:
	What is {Fill 3: <Name>'s/your} race? ({Fill 2: Is <Name>/ Are you} White; Black, African American, or Negro; American Indian or Alaska Native; Asian; Native Hawaiian or Other Pacific Islander; or Some other race?)
	<11> White <12> Black, African American, or Negro <13> American Indian or Alaska Native <14> Asian <15> Native Hawaiian or Other Pacific Islander <16> Some other race

Figure 3.2.3 – Specification Example with Variable Wording (continued)

Question Text

If not first or second time question is asked and previous person RACT NE 13, 14, 15 or 16:

How about {Fill 4: <Name>/you?}

{(Fill 2: Is <Name>/ Are you} White; Black, African American, or Negro; American Indian or Alaska Native; Asian; Native Hawaiian or Other Pacific Islander; or Some other race?}

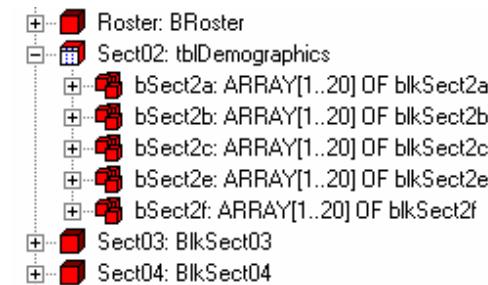
- <11> White
- <12> Black, African American, or Negro
- <13> American Indian or Alaska Native
- <14> Asian
- <15> Native Hawaiian or Other Pacific Islander
- <16> Some other race

3.3 Design Considerations and Implementation

The Content Test instrument would be developed using Blaise 4.6 build 965, to match the version with which the current 2008 ACS Production instrument was being developed.

In order to move through the table in a columnar, or topic-based, fashion, it was necessary to break up each question into its own table. The demographic block itself was defined as a table, with each question an arrayed block inside – two relationship question blocks (Sect2a1 and Sect2a2), a gender block (Sect2b), an age block (Sect2c), two Hispanic origin blocks (Sect2e1 and Sect2e2), and a race block (Sect2f). Only the gender block is purely topic based. It didn't make sense to break up the three fields that collected the month, day, and year, nor the age probe questions that followed them. Neither was it logical to move the "specify other" race questions apart from the leading race question. Several of the other question blocks also contained separate questions depending on whether a telephone or personal interview was being conducted. Eventually, the survey methodologists decided that they preferred to also collect the relationship questions and Hispanic origin questions as groups instead of individually, leaving us with five tables instead of the initial seven (see Figure 3.3.1).

Figure 3.3.1 – Data structure of the Demographic Section



The original person-based instrument contained all of the data in a single table. By breaking the topic-based instrument into pieces, the entirety of the demographic data was no longer available to the interviewer at any point during the demographic portion of the interview. To circumvent this problem, the programmers decided to display, or "show", all of the data that was not currently being "asked". In this way, the interviewer would

still progress down a single column without losing the ability to see all the data that had been collected thus far (see Figure 3.3.2).

Figure 3.3.2 – Screenshot of Table With All Data Displayed

Nonetheless, a straight “show” of the extra data was not ideal, as not all of the data would always be on-path to be collected and this would cause empty fields to be “shown”. This meant that rules were required even for showing the fields not collected in the current table. Rather than painstakingly coding the rules to “show” each individual field, the programmer placed the fields into blocks that mimicked the sub-blocks that were being “asked”, and performed the “show” method on the entire block, and thus all of the fields it contained (see Figure 3.3.3).

Figure 3.3.3 – Code from the Sect2a block demonstrating ASK vs. SHOW of sub-blocks

```

bRel_Questions ({ IMPORT} IN_lang_ext, LNO, IN_NAME, bSex_Display.SEX,
                 IN_HHolderVar, IN_HHNAME, IN_RESPLINE,
                 IN_MODE, IN_TQA_FLAG, IN_PV_TELE,
                 { TRANSIT} TR_REL, TR_RelNum)

bSex_Display.SHOW ({ IMPORT} IN_lang_ext, LNO, IN_NAME, IN_RESPLINE,
                     { TRANSIT} TR_SexNum)

bAge_Display.SHOW ({ IMPORT} IN_lang_ext, LNO, IN_NAME, bSex_Display.SEX,
                     IN_HHolderVar, IN_HHNAME, IN_RESPLINE,
                     { TRANSIT} TR_TempDOB)

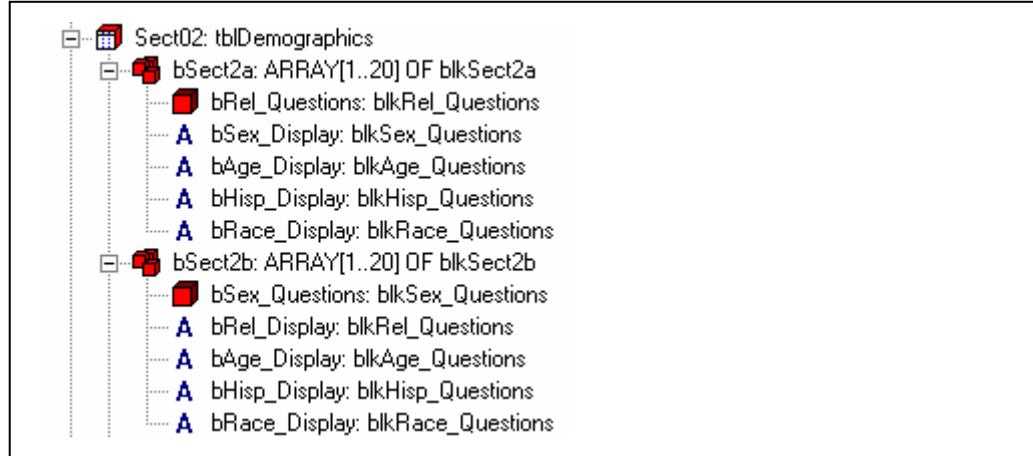
bHisP_Display.SHOW ({ IMPORT} IN_lang_ext, LNO, IN_NAME, bSex_Display.SEX,
                     IN_RESPLINE, IN_MODE, IN_TQA_FLAG, IN_PV_TELE,
                     { TRANSIT} TR_HISAFlag, TR_HISANum, TR_HISBNum,
                     TR_RaceIntroFill, TR_RaceIntroFillSP)

bRace_Display.SHOW ({ IMPORT} IN_lang_ext, LNO, IN_NAME, bSex_Display.SEX,
                     IN_RESPLINE, IN_MODE, IN_TQA_FLAG, IN_PV_TELE,
                     TR_RaceIntroFill, TR_RaceIntroFillSP,
                     { TRANSIT} TR_RACNum, TR_RACFlag, TR_RCW1Num, TR_RCWAGNum,
                     TR_RCWPNum)

```

With each table being basically five copies of the same table, but with a separate set of questions being “asked” versus “shown” in each, the programmer was going to have to maintain five sets of the same questions. To cut down on the amount of duplicated code, a block type was created for each question set that would be reused in each of the five tables, both those question blocks being “shown” and those being “asked” (see Figure 3.3.4). Because the block was defined in a single place, any changes made to it would apply to all tables. This improved efficiency, as well as long-term maintainability.

Figure 3.3.4 – Sub-blocks defined as fields or auxfields, of same block type



Another benefit of using the same blocks in each table was the ability to copy an entire block at once at the higher demographic level, making it very easy to keep all five tables up-to-date with the current data without having to pass large numbers of parameters from one to another and compromise performance (see Figure 3.3.5).

Figure 3.3.5 – Code from the Sect2a block demonstrating ASK vs. SHOW of sub-blocks

```

bSect2a[I].bSex_Display := bSect2b[I].bSex_Questions
bSect2a[I].bAge_Display := bSect2c[I].bAge_Questions
bSect2a[I].bHisp_Display := bSect2e[I].bHisp_Questions
bSect2a[I].bRace_Display := bSect2f[I].bRace_Questions
  
```

We found ourselves dealing with a demographic section five times the size of the original demographic section. To cut down on the size of the database, and minimize confusion with multiple copies of fields being output from the instrument, all sub-blocks that were being “shown” were defined as auxfields (see Figure 3.3.4). Therefore, in each table, only the sub-block that was being “asked” would actually be output. Also, the auxfields are not saved in the database, cutting down on the amount of time needed to save to the database upon closing the instrument.

The change to the demographic section had an impact on other parts of the instrument. Demographic data gets used throughout, for generating fills and determining question pathing. With the table now being split into five distinct tables, the full paths of fields changed. To reduce how many places the field name would need to be modified, explicit parameters were used as extensively as possible.

There were several obstacles that had to be overcome by using this particular design. Each block was being “asked” once, but “shown” an additional four times. When the

“show” method was applied, certain checks would be triggered again, meaning a single check might appear a total of five times. In some cases, where the check was conditioning on a calculated value, as opposed to a field within the block, none of the involved fields would be on path, and the instrument would get stuck. To circumvent this problem, a flag was added so that the block would know whether or not the “show” or “ask” method was being applied, and in the case of the “show”, the check would be ignored. In another situation, several checks had been embedded within the sub-blocks, but conditioned on data from several sub-blocks. By moving these checks up to the main demographic level, not only was the problem of the checks being triggered multiple times avoided, but also a number of generated and explicit parameters were eliminated.

3.4 Comparison to Topic-based Demographics in SIPP

In a previous project, the Survey of Income and Program Participation (SIPP), the authoring staff was required to program the demographics section of a Blaise instrument using a topic-based approach. Unlike ACS, however, the SIPP survey designers did not request to display the demographic data within a single table. Instead, each arrayed block was displayed separately. The existing SIPP instrument was programmed using the Computer Assisted Survey Execution System (CASES), and having only ever worked with SIPP as designed in CASES, the interviewers were accustomed to seeing only one field at a time. ACS, on the other hand, was already coded as a Blaise instrument, and had been fielded in production with the person-based mode, with all fields in one table, for several years. Therefore, to smooth the transition from person-based to topic-based, it was necessary to make the display change virtually transparent.

4. Conclusion and Recommendations

4.1 Flexibility Within the Table

While the ability to use free-navigation within tables in 4.7 has greatly increased the flexibility of data collection in Blaise instruments, it would be helpful to have the ability to design an instrument selecting the direction of flow through a table without sacrificing a structured interview where the interviewer continues to be guided from question to question.

In this example, where we had combined vertical and horizontal movement, it would probably still be necessary to create multiple arrayed blocks. All five arrayed sub-blocks were included in the same table. However, each sub-block was displayed on its own page, making it necessary to create copies to mimic a single table. The task would have been greatly simplified had Blaise had the ability to display all five arrayed sub-blocks on the same page. An additional drawback of the current implementation was that the interviewer could not use the mouse to skip backwards or forwards to edit a field only being displayed on the current page. It was necessary to use the arrow keys or page up and page down keys to navigate. While CAPI field interviewers are primarily accustomed to using the keyboard to navigate, should an interviewer choose to use the mouse, the fields being “shown” could cause confusion.

The way in which Blaise performs the “show” method on a block caused problems in this example. Checks were triggered even in the case where all fields involved in the check were themselves being “shown” instead of “asked”. It would be easier if checks within the block were treated the same way they were treated during a “keep” method.

4.2 Output Processing

For this particular project, the Blaise database was converted to a flat ASCII file for output processing. If the data were being output in an ASCII relational format, it is important to note that the five arrayed blocks would be output in separate files, and it would be helpful to have linking variables, such as the person line number, added to each sub-block to aid in piecing together a person's data.

5. Acknowledgements

5.1 Acknowledgements

Many individuals have worked and continue to work on components of the American Community Survey to ensure its successful operation. The authors would like to include grateful acknowledgement to all of our colleagues from the U.S. Census Bureau whose work directly contributed to the instrument development discussed in this paper. We thank in particular Li Li, Kenneth Stulik, Gianna Dusch, David Raglin, Courtney Reiser, Diane Cronkite, Todd Hughes, and Debra Klein for their work in the coordination, design, programming and testing of the instrument used in this phase of the project.

This document is to inform interested parties of ongoing research and to encourage discussion of Blaise instrument design issues. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

5.2 References

Moore, J. and Moyer, L. (2002), ACS/CATI Person-Based/Topic-Based Field Experiment, U.S. Census Bureau, Statistical Research Division, Research Report Series (Survey Methodology #2002-04), issued February 19, 2002